DOCUMENT RESUME

D 185 159	UD 020 353
UTHOR	Schwartz, Judah L.
ITLE	What Do We Know About Teaching and Learning in Urban Schools? Volume 3: Assessment that Respects Complexity in Individuals and Programs.
NSTITUTION	Central Midwestern Regional Educational Lab., Inc., St. Louis, Mo.
PONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
UB DATE	79
OTE	71p.; Paper prepared for the Urban Education Program, CEMREL, Inc.'s National Conference on Urban Education (St. Louis, MO, July 10-14, 1978). For related documents see UD 020 351-361 and UD 020 363.
DRS PRICE	MF01/PC03 Plus Postage.
ESCRIPTORS	Accountability; *Change; Change Strategies; Culture Fair Tests; Educational Assessment; Elementary Education; Institutional Evaluation; Program Evaluation; *Psychometrics; *Testing Problems; *Tests; Test Validity; *Urban Education
DENTIFIERS	*Project TORQUE

BSTRACT

Educational assessment tools are used for ccountability; selection and licensure, and to measure the effects f instruction for student diagnosis and treatment. Psychometric nstruments currently in use are flawed in two ways: they attempt to ank people on fundamentally multidimensional traits, and the problem f the validity of these instruments has never been resolved. Urban inority children have been particularly ill-served by the raditional paradigm of psychometric testing. The need for better ssessment techniques is urgent. Project TORQUE was initiated as a odel of alternative assessment. The goal was to design criterion eferenced tests of mathematical competence that are mathematically orrect, culturally fair, easily administered, and diagnostically nformative. Evidence of the validity of TORQUE instruments is erived from consistency of performance between test and validation xercises for each subskill. TORQUE materials are enjoyable and leasing for students and provide teachers with useful diagnostic uidelines. Alternative assessment techniques like TORQUE can ramatically expand the nature of what is known about a child's ntellectual growth. Further research is needed to answer questions bout assessment outside the mathematics realm and for purposes other han diagnosis and treatment. Educators can immediately alleviate ome problems of assessment by making public all tests following dministration and by forming groups to set specific goals for hildren, classes, and schools. (Author/MK)

****
Reproductions supplied by EDRS are the best that can be made *
from the original document. *
****

ED170113

WHAT DO WE KNOW ABOUT TEACHING AND LEARNING IN URBAN SCHOOLS?

Volume 3

Assessment That Respects Complexity In Individuals and Programs

Judah L. Schwartz
Massachusetts Institute of Technology

Paper prepared for the National Conference on Urban Education,
held in St. Louis, Missouri, July 10-14, 1978.
Conducted by the Urban Education Program, CEMREL, Inc.,
with support from the National Institute of Education.

CEMREL, Inc., St. Louis, Missouri

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Prepared by CEMREL, Inc., a private nonprofit corporation supported in part as an educational laboratory by funds from the National Institute of Education, Department of Health, Education and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement should be inferred.

Copyright on this document is claimed only during the period of development, test, and evaluation, unless additional authorization is granted by the National Institute of Education to claim copyright on the final version. For information on the status of the copyright claim, contact either the copyright proprietor or the National Institute of Education.

Copyright 1979, CEMREL, Inc.

Library of Congress catalog card number 79-54732
Manufactured in the United States of America

CEMREL, Inc.
3120 59th Street
St. Louis, MO 63139

Table of Contents

ACKNOWLEDGMENTS..... vii

FOREWORD..... xiii

ASSESSMENT THAT RESPECTS COMPLEXITY IN INDIVIDUALS AND PROGRAMS..... 1

I. BACKGROUND..... 1

II. THE EFFECTS OF INSTRUCTION -- FOUR QUESTIONS AND A FRAMEWORK FOR SEEKING ANSWERS..... 8

III. DETECTING THE EFFECTS OF INSTRUCTION..... 12

IV. WHEN SHOULD ONE LOOK FOR THE EFFECTS OF INSTRUCTION?..... 31

V. USING ALTERNATIVE ASSESSMENT APPROACHES, WHAT CAN ONE LEARN ABOUT CHILDREN?..... 35

VI. USING ALTERNATIVE ASSESSMENT APPROACHES, WHAT CAN ONE LEARN ABOUT PROGRAMS?..... 41

VII. SOME PROMISING RESEARCH DIRECTIONS..... 45

VIII. TWO POLICY RECOMMENDATIONS..... 48

REFERENCE NOTES..... 51

REFERENCES..... 52

Acknowledgments

The Urban Education Program, CEMREL, Inc., would like to thank the National Institute of Education and the Regional Joint Planning Group of the Urban Education Program for its support and assistance in planning the conference.

Conference Chairperson

Harriet Doss Willis
Director
Urban Education Program

Conference Participants

Nettie S. Armmer
Research Associate
St. Louis University

Eva Baker
Director
Center for the Study of Evaluation
UCLA Graduate School of Education

Carrie E. Bash
Division of Community Programs
Urban League of St. Louis

Isabel Beck
Co-director
Language & Communications Unit
Learning Research and Development Center
University of Pittsburgh

Janet Cheatham Bell
Urban/Ethnic Studies Coordinator
Indiana Department of Public Instruction

David Bennett
Deputy Superintendent
Milwaukee Public Schools

Edward Binkley
Director of Research & Evaluation
Metro Schools
Nashville, Tennessee

Milton Bins
Senior Associate
Policy Development
The Council of Great City Schools
Washington, D.C.

Dean Bowles
Director
Home-School-Community Project
Wisconsin Research and Development Center

Ronald A. Boyd
Associate Superintendent
Department of Education
State of Indiana

Emerson Brown
Board of Directors
CEMREL, Inc.

L.D. Buford
Principal
St. Louis Public Schools

William D. Callian, Jr.
Assistant Superintendent
Curriculum Instruction
Memphis City Schools

Angeline P. Caruso
Associate Superintendent
Department of Curriculum &
Instructional Services
Board of Education
City of Chicago

Courtney Cazden
School of Education
Harvard University

Ned Chalker
Teaching & Learning Program
National Institute of Education

Francis S. Chase
Director of Urban Education
Dallas, Texas

Rosemary Christensen
Director of Indian Education
Minneapolis Public Schools

Doris J. Clemons
Region Five Office
Detroit Board of Education
Detroit, Michigan

Frances Coe
Board of Directors
CEMREL, Inc.

David Colton
Graduate Institute of Education
Washington University
St. Louis, Missouri

Lloyd M. Cooke
Vice Chairman
Economic Development Council of
New York City

Dallas Daniels
Director
Equal Employment Opportunity
Indianapolis Department of Public
Instruction

A. W. Dirks
Director
Research, Planning & Development
Wichita Public Schools

Lorenza Dixon
Curriculum Coordinator
Indianapolis Public Schools

John W. Dobbs
Assistant Superintendent
School & Community Affairs
Department of Education
State of Michigan

Domingo Dominguez
Director
Bilingual & Migrant Program
Southwest Educational Development
Laboratory

Ronald Edmonds
Director
Center for Urban Studies
Graduate School of Education
Harvard University

Nancy Evers
Department of Educational
Administration
University of Cincinnati

Joseph L. Felix
Director
Program Evaluation
Cincinnati Public Schools

Marian Kilbane-Flash
Supervisor
Research & Development
Cleveland Public Schools

Anita Ford
Co-chairman
Citizens Education Task Force
St. Louis, Missouri

Ruges R. Freeman
Coordinator
Secondary Student Training
Southern Illinois University
Edwardsville, Illinois

Gary Gappert
New Jersey Department of Education

Booker Gardner
Wisconsin Research & Development
Center for Individualized Schooling
Madison, Wisconsin

Howard Gholson
Policy Analysis
Title I
U.S. Office of Education

Henry Givens
Assistant Commissioner
Department of Elementary &
Secondary Education
State of Missouri

Edmund Gordon
Director
Institute for Urban &
Minorities Education
Teachers College
Columbia University

Milton Goldberg
Educational Organizations &
Institutions
National Institute of Education

John H. Grate
Director
Planning & Development
Cincinnati Public Schools

Eldon Grossner
Manager
LEA Services
Illinois Office of Education

Rex Hagans
Northwest Regional Educational
Laboratory

Robert J. Havighurst
Department of Education
The University of Chicago

Garland M. Hawkins
Director
Music Education
Curriculum Center School Dist. 189
East St. Louis, Illinois

Paul Hill
The Rand Corporation

Keith Hyde
Planning & Development
Des Moines Public Schools

John Ingram
Wisconsin Research & Development
Center for Individualized Schooling

Barbara Jackson
Associate Dean
Department of Administration &
Supervision
School of Education
Atlanta University

James N. Jacobs
Superintendent
Cincinnati Public Schools

Sylvia K. Jones
State & Federal Grants
Toledo Public Schools

Preston C. Kronkosky
Field Services & Dissemination
Southwest Educational Development
Laboratory

Daniel U. Levine
Director
Center for the Study of
Metropolitan Problems
in Education
University of Missouri-Kansas City

Joseph D. Liggins
Assistant Superintendent for Staff Development
Houston Independent School District

Reece Little
Director of Curriculum Development
Jefferson County Public Schools
Louisville, Kentucky

Irene Lober
Superintendent
University City Public Schools
St. Louis, Missouri

William Lobosco
ESEA Title I
U.S. Office of Education

Catherine Lyon
Center for the Study of Evaluation
UCLA Graduate School of Education

Mary Meehan
District Coordinator
The School District of Kansas City

LaMar Miller
Director of Secondary Education
New York University

Len Nachman
Division of Planning & Development
Minnesota Department of Education

Tom Odneal
Director
Title IV
Missouri Department of Education

Joan Orender
Project Coordinator for Distribution of Information
Nebraska Department of Education

George A. Parry
George A. Parry & Associates
River Forest, Illinois

William Phillips
Minneapolis Public Schools

Andrew C. Porter
Institute for Research on Teaching
College of Education
Michigan State University

Electra Price
Oakland Unified School District
Oakland, California

Walter E. Richardson
Supervisor
Compensatory Instruction
Columbus Public Schools

J. Robert Sampson
Assistant Manager
Program Planning & Development
Illinois Office of Education

Judah L. Schwartz
Professor of Engineering, Science and Education
Massachusetts Institute of Technology

Cleona Shortridge
St. Louis Public Schools

Barbara Shull
Conference on Education
St. Louis, Missouri

Edward E. Smith
Right to Read Effort
Washington, D.C.

Douglas Smith
Administrative Assistant to the Governor
State of Michigan

Frank Sobol
Basic Skills Group
National Institute of Education

George Spicely
U.S. Office of Education

Jane Stallings
Stanford Research Institute

Tom Stefonek
Director of Planning & Evaluation Research
Wisconsin Department of Public Instruction

Vera John-Steiner
Department of Educational Foundations
College of Education
The University of New Mexico

Ronald Stodghill
Associate Superintendent
St. Louis Public Schools

Marilyn Suydam
ERIC Information Analysis Center for Science, Math & Environmental Education
The Ohio State University

Maud Thomas
U.S. Office of Education
Kansas City Regional Office

Concepcion Valadez
Graduate School of Education
University of California
Los Angeles

Mamie Wandick
Program Officer
U.S. Office of Education
Kansas City Regional Office

Phyllis Weaver
School of Education
Harvard University

Billy-Belle Weber
Weber, Fuhrmann & Associates
East St. Louis, Illinois

Joyce Weddington
Director
Research and Evaluation
Memphis City Schools

Rachel T. Weddington
University Dean for Teacher Education
The City University of New York

Lores Wells
Harris Teachers College
St. Louis, Missouri

Robert Wentz
Superintendent
St. Louis Public Schools

Barbara Woods
Afro-American Studies
St. Louis University

Saul Yanofsky
National Institute of Education

Savannah M. Young
St. Louis Public Schools

Allen F. Zondlak
Director of Planning
Detroit Public Schools

URBAN EDUCATION PROGRAM STAFF

John Bass
Resource Coordinator

Alfreda Brown
Resource Specialist

Pamelia Coe
Research Associate

Eileen Hamilton
Secretary

Nellie H. Harrison
Dissemination-Curriculum Specialist

Donald Miller
Research Advisor

Harriet Doss Willis
Director

Astacia Wright
Research Assistant

STAFF FOR WHAT DO WE KNOW ABOUT TEACHING AND LEARNING IN URBAN SCHOOLS?

Harriet Doss Willis
Editor

Margaret Solomon
Manuscript Editor
Production Coordinator

Linda Cambell
Office Manager

Marlene Dandridge
Floor Clerk

Arneita Gray
Secretary

Angela Reed
Secretary

Dian Taylor
Secretary
Office Manager

Karen Temmen
Secretary

Foreword

This monograph is one of a series of papers on the issues of instruction and learning in urban schools that were presented at a conference held July 10-14, 1978, in St. Louis, Missouri, sponsored by the Urban Education Program, CEMREL, Inc., and supported by the National Institute of Education. It is our hope that both the conference and the series of fourteen monographs that resulted from it will assist educational researchers and school practitioners in identifying and analyzing instructional and learning problems in urban school settings, and will contribute to the development of strategies for the improvement of schooling for students.

Titled "What Do We Know About Teaching and Learning in Urban Schools?", the 1978 conference focused on an examination of research findings on the teaching-learning process in areas most germaine to and under control of the schools. The key problems to which the papers are addressed are the assessment of learning outcomes and the analysis of the relationships between instructional and other inputs and learning outcomes. Important contextual issues not under the full control of schools, such as community and parental involvement and the impact of federal and state governments in improving educational opportunities, were addressed at the conference as well.

The conference was conceived, in part, as an initial step in the development of a state-of-the-art review of the most critical issues faced by educators in urban schools who struggle with the morass of general educational problems, exacerbated by decreasing public confidence in large city schools, declining student enrollment, and increasing numbers of poor and minority students. While the conference does not exhaust discussion of the problems

that are particularly significant in the urban school setting, it forms a convenient focus for viewing the larger issues affecting education in our cities.

The educational innovation and change growing out of more than ten years of effort since the passage of the Elementary and Secondary Education Act of 1965 -- the primary stimulus for federal support of research and development in education -- has had the least impact on the student populations typical of the urban school setting. With this concern as a primary consideration, ten months prior to convening the conference, CEMREL's Urban Education Program had begun work with a network of twenty big-city school districts and state departments of education in Illinois, Indiana, Iowa, Kansas, Kentucky, Michigan, Minnesota, Missouri, Nebraska, Ohio, Tennessee, and Wisconsin. Thus, the conference and the research findings of national authorities was undertaken to lay a foundation for future work in urban educational research and development and to undergird present efforts at improvement of schools in the region.

The Conference Design

To guide the development of the presenters' preparation of their papers, topical questions were prepared for them by Urban Education Program Staff and a paper format was recommended. Each presenter prepared a full paper based on the topical questions and results were summarized by the writers during general sessions of the conference. Because of a wide range of perspectives held on the conference issues, discussions precipitated by reactor panels composed of other prominent researchers, practitioners, and community leaders in education followed each general conference session. More than one hundred educational researchers and practitioners from across the country attended the conference. The invited participants came from school districts and state departments of education from nineteen states; thirteen institutions of higher education; four federal agencies; numerous educational laboratories and research and development centers; and seven social service agencies.

The conference was divided into two types of sessions: general sessions, wherein the nationally recognized speakers who wrote papers presented their summaries; and forum sessions, wherein reactors helped analyze the presentations and examine the concrete recommendations put forth by the general-session speakers. All participants were encouraged to present questions and reactions during these forum sessions.

Francis S. Chase, Director of Urban Education Studies, Dallas, Texas, gave the keynote address at the opening session of the conference, a banquet on Monday evening, July 10. His topic, "Promising Developments in Urban Education," summarized field studies that he was conducting under sponsorship of the Council of Great Cities Schools and the University Council for Educational Administration under a grant from the Spencer Foundation. Chase is Professor Emeritus of the University of Chicago, where he also served as Chairman and Dean of Education for ten years.

The Conference Topics

Each of the presenters had been asked to design a paper by responding to salient topical questions, drawn up by CEMREL's Urban Education Staff. The questions that were posed originally for each of the writers are presented below. Most of the resulting papers were developed into monographs for this series.

Compensatory Education: A Congressional Perspective Christopher T. Cross Minority Staff Director Committee on Education and Labor U.S. House of Representatives.	What is the current state of the evidence concerning projected changes in legislative mandates, and what may we expect related to changes in policy governing federally funded compensatory education programs?

Effective Teacher-Support Systems to Improve Instructional Conditions in Schools Joseph D. Liggins Assistant Superintendent for Staff Development Houston Independent School District	What teacher support systems are most effective in improving instructional conditions for students? What is the role of in-service training, and what types of in-service training can be recommended to urban educators? What other resources--such as additional curricular materials, support from teacher aides and volunteers, time re-allocations, team teaching arrangements, etc.--can be demonstrated to be effective in improving students' learning opportunities? What is the state of the evidence?
Assessment That Respects Complexity in Individuals and Programs Judah L. Schwartz Professor of Engineering, Science and Education Massachusetts Institute of Technology	How can we best measure and analyze the effects of instruction on students' learning in the major school subjects: What are the most appropriate times and strategies for measuring the effects of schooling on student learning? What do we learn from means of assessment other than standardized achievement testing? When and under what circumstances is it most appropriate to use alternative or complementary strategies for assessing program effectiveness?
Achievement Tests in Urban Schools: New Numbers Eva Baker Director Center for the Study of Evaluation University of California, Los Angeles	How can we best measure and analyze the effects of instruction on students' learning in the major school subjects? What are the most appropriate times to measure the effects of schooling on student learning? What are the cautions which should be taken into account before any attempt to evaluate the effectiveness of programs primarily on the basis on standardized achievement data, especially when the number of individual implementations is large and the actual conditions under which the implementations took place are not well documented?
Summer Drop-off and The Effectiveness of Compensatory Instruction Paul T. Hill The Rand Corporation	What has the era of compensatory education program development contributed to knowledge about achievement in the basic school subjects in urban schools? Where is the evidence? What, if anything, does the evidence indicate about methods for improving basic skills achievement? What, if any, recommendations can be made to school district planners on the basis of existing evidence?

Instructional Ingredients for the Development of Beginning Reading Competence Isabel Beck Co-Director Learning Research and Development Center University of Pittsburgh	Much research, program development, and implementation of efforts to improve achievement in basic school subjects has been concentrated in the area of reading. What approaches to instruction in reading have been successful in increasing achievement? Is there evidence that certain approaches are best suited to the needs of particular populations? Do these effects generalize to other areas of school learning, such as mathematics, social studies, and science?
A Discussion of The Literature and Issues Related to Effective Schooling Ronald Edmonds Director Center for Urban Studies Graduate School of Education Harvard University	Are there schools that have demonstrated a higher degree of effectiveness in the delivery of instruction in the basic school subjects than others serving similar student populations (in terms of socioeconomic backgrounds, family characteristics, minority populations, etc.)? What characteristics distinguish these schools from less successful schools in similar situations? What is the evidence that the explanatory characteristics have been correctly identified, and how strong is that evidence?
The Educational Promise of Cultural Pluralism Vera John-Steiner and Larry Smith Department of Educational Foundations College of Education The University of New Mexico	How can we approach problems of teaching and learning in urban schools so that the racial and cultural diversity of the school population is recognized, respected, and utilized constructively, to the extent possible? How do we differentiate between basic academic skills, which all students need to acquire for successful participation in our hetergeneous society, and curriculum decisions which can and should be tailored to the needs and interests of different sub-cultural groups? Do students benefit from participation in racially and culturally hetero-geneous classrooms, or perhaps from certain kinds of heterogeneous instructional settings but not others? What is known about these issues, and how much consensus is there about the cultural values on which schooling must be based?

New Perspectives on School District Research and Evaluation Catherine Lyon Center for the Study of Evaluation University of California, Los Angeles	What is the role of research and development processes in improvement of instruction and learning in basic school subjects in large urban systems? In what form are these processes being applied to the teaching and learning process in urban schools? What is the evidence and how solid are the indications that use of these proceses has resulted in improved student achievement?
Effective Teaching and Learning in Urban Schools Jane A. Stallings and Shirley W. Hentzell Stanford Research Institute	Looking specifically at the interaction of student, teacher, curricular and school organization characteristics, what instructional conditions encourage achievement in basic school subjects? What evidence is there, and how good is it? How are the effects of instructional conditions on student achievement measured? On the basis of existing evidence and trends, what recommendations can be made for improving students' instructional environments?
Mathematics and The Urban Child Marilynn Suydam Center for Science, Mathematics, and Environmental Education The Ohio State University	Mathematics is a school subject which is generally considered basic, insofar as mathematical skills are required for successful adult functioning in our complex society. What evidence is there to indicate which approaches to instruction in mathematics have been successful in increasing achievement? Is there evidence that certain approaches are best suited to the needs of particular populations? Do these effects generalize to other areas of school learning, such as reading, social studies, and science?
Basic Skills in Urban Schools: A View From The Bilingual Classroom Concepcion Valadez Graduate School of Education University of California, Los Angeles	Since the second largest minority population served by many urban school districts is a bilingual population, what is the current state of the evidence about the effects of programs developed especially for bilingual students? What, if anything, does the evidence indicate about the effects of such programs on basic skills achievement?

School-Community Relations and The Urban School Dean Bowles Director Home-School-Community Project Wisconsin Research and Development Center for Cognitive Learning	What do we know or what can we learn about the influence of parental and community involvement in schooling on student achievement? How does parental and community involvement affect the teaching-learning process? If information is not available to answer these questions on even a tentative basis, what types of research are recommended to study this aspect of students' lives more effectively?
Frank Sobol Basic Skills Group National Institute of Education	Are there research and development products which have contributed to improved instruction and increased student achievement in basic school subjects in large urban school systems? What is the evident and how solid are the indications that use of these products has resulted in improved student achivement?
A Summary of The Conference Findings Edmund W. Gordon Director Institute for Urban and Minority Education Teachers College Columbia University	In light of the evidence produced by examination of these issues, what concrete recommendations can be made to school districts concerning the improvement of instruction and learning in basic school subjects? In light of the same evidence, what research and development agenda should be the focus of future study and program development?
Dr. Francis S. Chase, Director of Urban Education Dallas, Texas	Keynote Speaker for Conference Banquet: "Promising Developments in Urban Education"

--Harriet Doss Willis
Director
Urban Education Program

Assessment That Respects Complexity In Individuals and Programs

Judah L. Schwartz
Massachusetts Institute of Technology

I. BACKGROUND

The psychometric tradition began to develop with some strength around the turn of the century. The science of psychometrics, defined as "the measurement of mental traits, abilities, and processes," had many rather diverse and enthusiastic early supporters. Perhaps the man who was most influential in establishing psychometrics as a discipline was Edward Thorndike, a student of William James. Thorndike was very taken with the success of quantitative methods in the natural sciences and, in fact, believed that the establishment of psychology as a science depended on its ability to make measurements in the mode of the natural science. Thorndike said, "Whatever exists at all, exists in some amount."

The emerging field of psychometrics was, in many ways, philosophically quite compatible with the contemporary views of social Darwinism. In fact,

---

JUDAH L. SCHWARTZ is Professor of Engineering Science and Education in the Division for Study and Research in Education and in the School of Engineering at the Massachusetts Institute of Technology.

the first products of that tradition--the early intelligence and achievement tests--did show that the performance of those in lower social and economic classes was inferior to that of people in higher social classes.

Testing and social class is an important issue in looking at the early history of testing in this country, because it is a prism through which one can view the two major early uses of tests, i.e., testing for accountability and testing for selection. Particularly in the large and growing urban areas of this country, the schools of the last 20 years of the nineteenth century and the first 20 years of the present century were a political battleground. Wave upon wave of immigrant groups fought for political ascendancy and power within this one institution, which seemed to offer a multitude of opportunities for control of jobs and, in a deep sense, for control of the culture of the still developing society. Achievement testing entered that arena in the service of those who tried to argue that the schools had not deteriorated now that they had fallen into the hands of the newly immigrated. By the First World War, achievement tests were in widespread use and began to sell in large numbers of hundred thousands per year.

At about the same time, the first major use of psychometric instruments for selection came into being. The administering of group intelligence tests by U.S. Army psychologists during the First World War was a procedure for the selection of officers and the screening of recruits felt to be inadequate for service. The apparent success of psychometric instruments in aiding this process led, after the war, to their establishment as legitimate tools to be used in the selection process.

That the standardized test was successfully established in the early part of this century as a tool for answering vital social needs is unquestioned. Since then, our society has developed and expanded and become substantially more open. Our purposes and needs for assessment have similarly expanded and become more diverse. The successes of the early part of the century need to be re-examined in light of what our society has become and is likely to become.

The Uses of Educational Assessment

Assessment for the purposes of educational accountability continues to be a major use of testing in our society. There has been increasing pressure at the federal level and in a number of states to introduce various forms of competency examinations and other accountability mechanisms. The response of the education establishment to these pressures is not yet clear. Nor is it clear what the consequences of a nation-wide set of minimum competency examinations are likely to be. The most enthusiastic proponents of such measures would like to see (on the evidence of their own testimony) such instruments developed to be used at every grade level throughout the country.

Testimony from respected members of the public interested in education and legislation introduced into Congress are not the only sources of pressure for increased accountability using standardized test instruments. The State of Michigan has made judgments as to the success of its educational system using such means for some time now. So has the City of Denver. In general, the use of psychometric instruments as the basis for judgment as to the success of school systems is increasing.

Psychometric instruments are not only used for accountability judgments about the quality of school systems. They are also used in order to make accountability judgments of the performance of individual teachers. This document is not the place to argue the merits or demerits of that particular use of tests. They have certainly been amply analyzed by both the National Educational Association and the American Federation of Teachers, and relevant arguments on both sides of the issue of using standardized tests to make accountability judgments about teachers can be found in the writings of those organizations.

The uses of tests for accountability purposes covers even wider territory. In Kalamazoo, Michigan, the salaries of teachers and administrative personnel in the school system are tied to the performance of individual children on psychometric instruments. In many places the allocation of

funds to school districts is tied to performance on such instruments. Frequently, the availability of funds for special title programs is tied by law to the performance on psychometric instruments.

Perhaps the most telling indication of all about the widespread use of psychometric instruments is the national concern for the drop in Scholastic Achievement Test scores that has attracted public attention in the last several years. The careful deliberations of the committee chaired by Willard Wirtz point to many possible reasons for this observed decline. Without attempting to discuss here the merits of that report or the merits of the instruments, the results of which the report investigates, I simply point to the fact that the results are taken as data regarding a real phenomenon, i.e., the drop in performance is taken as an indication of a lessened capability on the part of our graduating high-school students. Clearly, we, as a society, believe that we need these tests in order to have some indication of the quality of the high schools of this country and to hold educational authorities accountable for quality that does not match our expectations.

In addition to using tests for purposes of accountability, we also use psychometric instruments for the social purposes of selection and licensure. The frequent occurrence of names like Bakke, DeFunis, Griggs vs. Duke Power make it clear that these issues are of central importance to our society. In a very real sense, selection and licensure are two sides of the same coin. Selection tests determine who may have access to limited resources. Licensure tests determine who may have access to unlimited resource. In the educational context, selection tests are pre-treatment and licensure tests are post-treatment. Because selection tests are pre-treatment they typically determine necessary conditions for success in the subsequent educational situation. Because licensure tests are post-treatment they typically determine sufficient conditions for success in the subsequent situation.

Finally, and perhaps most importantly, we use tests to detect and assess the effects of instruction for the purposes of diagnosis and treatment, in the broadest sense. We would like to be able to use the results of tests--and often publishers tell us we may do so--in order to help shape and inform instruction. They say that we may infer the strengths and weaknesses of a student's performance from his or her answers on the tests, and use those data as a way of guiding future instructional strategies with that student.

There are questions that need to be raised so that we may understand how psychometric instruments can play a diagnostic role in education. Psychometric instruments are almost never based on an explicit underlying theoretical cognitive model of the tasks, and one might seriously question the utility of any diagnoses derived from the resulting answers to the questions on the test. Psychometric instruments are frequently normed against representative populations and thereby allow substantial built-in cultural, linguistic, and ethnic bias to go undetected for long periods of time. Even in those instances in which explicit efforts are made to norm tests against varying kinds of groups of children or students, what procedures have we evolved to allow the test taker to decide the appropriate norm population that he or she should be compared with? Further, as the Wirtz committee so ably points out, in the past decade our schools have seen new populations of students with new sets of aspirations. Are the instruments we now have, or the instruments that we are capable of generating within the psychometric paradigm, capable of dealing with these new students and their aspirations?

It is not our purpose here to present a detailed analysis of the social uses of assessment. Suffice it to say that accountability, diagnosis, selection, and licensure are all legitimate, indeed, necessary functions in a society. In fulfilling these functions we have come to rely heavily on psychometric instruments. But there are reasons to question how well psychometric instruments can allow us to fill these societal functions in an enlightened and human way. Indeed, there is reason to question whether or not we are

fulfilling these functions in an efficient way using these instruments. In particular, psychometric instruments are flawed in such a way as to make their utility for assessing the effects of instruction virtually negligible. We shall analyze very briefly the nature of two of these flaws and then turn our attention to the question of how best to devise and implement alternative assessment strategies.

The Methodological Inadequacies of the Psychometric Paradigm

There are two broad classes of objection that can be raised to the dominant paradigm in use in the psychometric community. The first of these is an objection to instruments that rank order people on fundamentally multidimensional "traits, abilities, or behaviors." The second objection is of a different character, i.e., psychometric instruments are almost always validated in derivative ways, and the problem of their validity has never been resolved.

We have known since ancient times that well ordering is a property of one-dimensional attributes. Multidimensional phenomena of all sorts cannot be well ordered except when a theory of aggregation is present. Under those circumstances, it is possible to aggregate complicated and diverse phenomena so that they differ from one another only with respect to the value of the attribute in question. In the absence of such a theory, in the absence of understanding the nature of the aggregation that one must perform in order to rank order the values of the single attribute in question, in the absence of an understanding of the nature of the attribute itself, and in the absence of the ability to detect the amount of attribute that may be present in any circumstance, it is, of course, not possible to rank order the value of the attribute (Schwartz, Note 1).

The problem of validity is the second major, very serious issue that one must raise in looking critically at the psychometric instruments in current use. A comprehensive review of the literature on validity in all its forms

(see Cronbach, 1970) is not appropriate here. Suffice it to say that the validity of achievement tests as measures of educational accomplishment is rarely questioned (Levine, 1976). In the area of criterion-reference testing, little work has been done on establishing with clarity a possible meaning or meanings for validity (Stake, Note 2). The general methodological problem is clearly far from settled. Thoughtful discussions (Messick, 1975) continue to appear with regularity.

The problem of the unease with traditional notions of validity is not a new one. Recently, Oscar Buros(1977), the late director of the Institute of Mental Measurements and editor of the Mental Measurements Yearbook, quoted his own criticism of validation procedures written in 1935. By and large Buros felt that the procedures in common use in 1977 were sufficiently unchanged to allow his 1935 criticisms to stand. It is illuminating to quote him. In commenting on choosing items for their discriminating ability, rather than for their quality of content, he says:

> It seems inescapable that such methods of statistically validating achievement tests insidiously tend to strengthen the status quo, to impede curricular progress, to perpetuate our present grade classification, to differentiate rather than to measure, to conceal unlearning and to give an illusory sense of continuous learning from grade to grade.

He goes on to say:

> If we make it our goal to measure rather than to differentiate, most of our methods of constructing tests, measuring repeatability, assessing validity, and interpreting test results will need to be drastically changed. (pp. 9-16)

Psychometrics as a strategy for evolving instruments to be used for the necessary and vital social purposes of assessment thus finds itself between a rock and a hard place. There is growing, thoughtful criticism of the tradition of psychometrics on both of these counts. The problems lie with the impossiblity of rank ordering on the gounds of the multidimensionality of the phenomena and with the deep methodological troubles with establishing validity. Pressures for evolving new and more acceptable paradigms are growing, and properly so. Critical analyses of the flaws of existing paradigms, while necessary, are hardly sufficient. If the field of education is to make progress and to go beyond the instruments that may have once served it well but no longer do so, it must look forward to the time when new strategies and new instruments that will serve evolving social goals can be developed.

II. THE EFFECTS OF INSTRUCTION--FOUR QUESTIONS AND A FRAME-WORK FOR SEEKING ANSWERS

We now turn our attention to a set of interrelated questions, each of which deals with an aspect of detecting the effects of instruction and assessing the quality of those effects, either in terms of the learning of individual children or in terms of the effectiveness of programs. The first of these questions is:

> How can we best measure and analyze the effects of instruction on students' learning in the major school subjects?

Clearly, the concern for this issue does not exist in a vacuum. That concern immediately prompts the next question, which seems at first to deal only with the operational dimensions of assessing the effects of instruction, i.e.:

> What are the most appropriate times and strategies for measuring the effects of schooling on student learning?

Undoubtedly, there are many who are satisfied with existing instruments as vehicles for providing answers to these questions. But these traditional vehicles have come increasingly under attack. Others at this conference will address the question, What can be learned from standardized achievement testing as an assessment strategy? Here, we turn out attention to a somewhat different question:

> What do we learn from means of assessment other than standardized achievement testing?

It has been argued elsewhere (Houts, 1977) that nothing of use or of interest is learned from norm-referenced achievement testing. Nonetheless, for purposes of diagnosis, prescription, and accountability, it is crucial that we learn about the effects of instruction in our schools. I take the view that there are many things that can be learned, but they can be learned only through means of assessment that do not fall under the traditional rubric of standardized norm-referenced testing. And finally, we are interested in the question:

> When, and under what circumstances is it most appropriate to use alternative or complementary strategies for assessing program effectiveness?

I shall try, in the forthcoming sections of this essay, to address these questions with particular emphasis on children in urban schools. Urban children, poor children, minority children have all been particularly ill-served by the traditional paradigm of psychometric testing, and the need to provide the children, their parents, and their teachers with better techniques and instruments of assessment is, indeed, urgent.

My analysis of these questions will be strongly shaped by my view of the nature of allowable inference and the logical consequences of that view applied to the problem of assessment. It is well, therefore, to expose my own perspectives explicity.

When we pose a question to a child in order to ascertain whether or not the child can perform a certain task, we ordinarily make the assumption that if the child does perform the task, he or she can perform the task, but if the child does not perform the task, he or she cannot perform the task. There is clearly a logical difficulty. Competency, as a construct, is not observable. Performance is. Inevitably we are obliged to prove competence via observed performance. We make inferences about competence from our observations of performance. What sorts of inferences are legitimate?

It is logically unacceptable to simultaneously hold the view that performance implies competence and that non-performance implies non-competence. If we accept the view that performance implies competence, we can make no inference whatsoever about competence in the case of non-performance.

This asymmetry between the kinds of inferences that one can make in the event of performance and in the event of non-performance is a fundamental one. It must not be neglected, for it has consequences of very great importance to the entire problem of assessment.

Because of this asymmetry, it seems to me that if we observe an instance of a child not performing when we probe some skill, we have an obligation to the question, perhaps in another context, or another form, or another format, or another style, so as to elicit performance from the child.

The problem of reformulating the probe, however, is a delicate one, because it calls for a level of understanding of the subtlety of both the task and the context that, in almost all cases, we do not possess. As our knowledge of cognitive development grows, we stand more and more in awe of the complexity of even the most apparently simple tasks.

Above and beyond our obligation to try and reformulate our probes of children's competence in ways that reflect the complexity of the tasks, we have a further obligation to recognize the idiosyncratic variation of

children's backgrounds and experiences. No child in school is a tabula rasa. They have all had a wealth of diverse experience both outside school and within it. Assessment strategies must, in the end, learn how to be sensitive to that dimension of the child.

In summary, therefore, the asymmetry of allowed inference in the assessment situation leads inevitably to the obligation of reformulating probes that are used for assessment. The obligation of reformulation leads inevitably to the need for subtler understanding of the tasks and the contexts and the idiosyncratic background of the children. These perspectives will all be called into play explicitly when we address the questions that were posed at the beginning of this section.

It is important before I proceed, however, to make clear a belief of mine. I do not believe that the intellectual activity of people can be measured. I do not believe that people can be regarded as the working fluid of an engine. I do not believe that individual human beings are statistically interchangeable. I do not believe that urban children or suburban children form homogeneous populations. I believe people come in groups of one.

In all but the most trivial classification schemes, the number of pertinent, important, and relevant attributes that characterize any human activity are so great that the likelihood of finding two people who are substantially the same is small. In particular, if the domain of interest is what and how students learn in schools, then it is only with the most simplistic reductionism that one can talk about high socioeconomic levels, low socioeconomic levels, black children, white children, Latino children, male children, female children, children of particular ages, and similar categories that generate partitions of the populations.

Finally, I do not believe the problem we face is how to find a version of traditional assessment practices and instruments that can be applied in urban settings without injuring the children. On the contrary, I see our problem to be the need for new assessment paradigms and procedures for all children. As has happened frequently in the last two decades, the victims of bias and prejudice are forcing the society as a whole to inspect its rhetoric and its institutions. No child in America is well served by assessment instruments built on an intellectually bankrupt foundation. The urgency of the problems of urban education can force us all to recognize that and continue to improve education for everyone.

III. DETECTING THE EFFECTS OF INSTRUCTION

Project TORQUE--A Model for Alternative Assessment

I turn now to the first of the four questions I was asked to address when I was invited to write this paper:

> How can we best measure and analyze the effects of instruction on students' learning in the major school subjects?

I wish to amend this question in several important respects. To begin with, the word best should be replaced by better. It is unlikely that we will ever devise an optimum solution to any educational problem. Secondly, I wish to replace the word measure by assess. I do believe, and have argued in detail elsewhere (Schwartz, Note 1), that the word measurement is appropriately used in evaluation contexts. Achievement is not a measurable magnitude, although it may be assessable. This is not a narrow scholastic quibble. The use of measurement implies a precision in the definition of the construct to be quantified and is accompanied by an expectation of accuracy that is clearly irrelevant in educational contexts. Thirdly, I wish to replace the words the major school subjects by the words language and mathematics.

School subjects can be broadly characterized as either tapestry subjects or vehicle subjects. Language and mathematics, the two great symbol systems that the human race has created, constitute the vehicle subjects of our schools. Science, literature, art, social science, history, and music are among the tapestry subjects. To a first approximation it is true that for most people, the vehicle subjects have no intrinsic content. Our reading, writing, speaking, and listening all refer to people or places or things. Mathematics is present in the schools because it provides us with the tools of analysis to model complex situations that have quantitative aspects to them. Our notions of competence and our expectations of achievement are rather different in the vehicle subjects and the tapestry subjects, whether we are focusing on student achievement on the one hand, or program effectiveness on the other. Here, in discussing student achievement I choose to limit the discussion to the vehicle subjects.

In order to devise better assessment instruments than we now have to assess students' learning, it seems to me that a number of criteria must be met:

1. Better assessment instruments must be flexible and must be able to accommodate the idiosyncratic learning styles of different children.
2. Assessment instruments in principle cannot be free of specific cultural or ethnic or linguistic assumptions. Better assessment instruments should make clear and explicit the assumptions they do make. This is particularly important if one wants to see the effects of instruction as distinct from the child's prior knowledge.
3. Of necessity, all assessment instruments are probes of the child's knowledge and of the child's comprehension of the language that is used to pose the question. It is in principle impossible to separate the two. It is possible, however, to reduce the confounding of one by the other. Better assessment instruments should do so.

4. The technology of providing answers should be adjusted so that it is comfortable for the child who is providing the answers and not for the computer that is grading the papers.
5. Better assessment instruments should have diagnostic value. Moreover, the diagnostic information they provide should be made available quickly to students as well as to teachers.
6. There is no justification for assessment instruments that cannot be made public in their entirety. As a society, we endure many costs for assessment devices that are not made public. Among them are cultural and ethnic biases that can only be discovered long after the damage they have done has been perpetrated, and then only by statistical analyses. Another cost of secret tests is the generally pedestrian level of the intellectual quality of the items. Without public exposure of the tests, there can be no public debate about the degree to which these instruments contribute to the raising of quality in the schools.
7. Better assessment instruments must be such that they engage the students' full efforts without threat and with minimum anxiety.

This is not an unreasonable list of criteria that we insist better assessment instruments should meet. People who have thought about the problem of alternative assessment have long recognized the importance of these criteria. In many respects, so has the larger educational community. But the use of traditional, standardized norm-referenced instruments continues with the claim that nothing better exists.

I propose to present a case study of an alternative strategy for developing and validating asessment instruments that a) exists; and b) has produced assessment instruments that satisfy these criteria.

Starting in the spring of 1974, Project TORQUE set out to design criterion-referenced tests of mathematical competence for elementary-school children. The project undertook to design tests that were mathematically correct, culturally fair, easily administered, diagnostically useful for teachers and students, and informative for administrators and parents. The project insisted that the validation of the tests depend on observations of actual behavior in a variety of tasks that children find attractive, including games and hands-on activities, rather than on other tests or on expert opinion. As a matter of principle, the project works with about as many minority children as majority children.

Computation tests that satisfy these criteria have been developed with the aid of approximately 10 thousand children and their teachers around the country. In addition, tests of length, area, and time measurement have been developed and validated. A great deal of clinical experience with games and other validating activities has been accumulated with children of diverse cultural and linguistic backgrounds, in such areas as the measurement of volume, weight, temperature, angle, and the skills of estimation, mapping and graphing, and word problems. Estimation and solving word problems form the heart of the application of mathematics to real life. Both are complex skills to analyze and difficulty competences to assess with pencil and paper tests. Progress in cognitive psychology in general and the reserarch results of Project TORQUE and groups like it in particular are making it possible for the first time to assess competence in these skills in ways that are useful and meaningful.

In addition to developing new assessment instruments and strategies, Project TORQUE is exploring with teachers, parents, and administrators the processes by which local consensus can be reached on what skills should be taught in the elementary schools and how school performance in teaching these skills may be assessed.

On the surface, Project TORQUE seems to be an effort addressed to the development of assessment in mathematics, but it is a good deal more than that. It has broken new methodological ground in the area of validation. It has devised assessment instruments that meet all of the criteria enumerated above for better assessment instruments. Further, as its work progresses, the generalizability of its methods to assessment of language competence becomes increasingly clear. For all these reasons, therefore, I wish to look at three aspects of Project TORQUE's approach to assessment. These are the validation of assessment instruments, the assessment instruments themselves, and the setting of performance criteria.

A New Approach to Validity

The key to TORQUE's validation strategy is the comparison of the performance of children on the test to be validated with their performance on specially devised games and other activities. This comparison is made for each of the subskills of the task, competence on which the test is trying to assess. Evidence that the test is valid derives from the consistency of performance between the test and the validation activities for each subskill.

The validation method grows naturally from TORQUE's procedure of developing the tests that are to be validated. This procedure is different enough from those used for traditional standardized tests so that it may be helpful to outline, from beginning to end, the history of the development of one test--that for the measurement of length.

The first step was to discuss length measurement with classroom teachers and as a staff. Based on these discussions, TORQUE developed more than a half-dozen games and activities that required children to measure lengths of both drawn objects (e.g., lines, sides of triangles) and real objects (e.g., length of a table, width of child's own handspan).

The project used games for two reasons: First, it is one way to get maximal attention and effort from elementary school children--a necessary requirement if their ability to measure is to be sampled reliably; second, the game is one step removed from "real measurements" and represents an idealization that helped to develop a general model for the process people use in making measurements.

Project TORQUE tried out the games and activities with small numbers of children (usually on the order of a dozen or so) and then with larger groups. The graphic presentation was refined and the resulting games printed in multiple copies. They were then tried out in classrooms by teachers who had helped to develop the games and by others who were extensively briefed. Their suggestions and the project staff's developing insights were embodied in further revisions, which were then tried out by a larger number of teachers who were less centrally involved in the project and so provided an "outside" judgment of the games. Typically, this latter stage involved 35 teachers and 1,000 children.

Parallel with the development of games and activities and informed by this development, was the evolution of a simple general model of the measurement process. Most physical measurements (of extensive magnitudes) involve the following steps:

1. Identifying the attribute to be measured (e.g., <u>length</u> of the <u>longest side</u> of <u>which triangle</u>).
2. Choosing an appropriate unit (judging "appropriateness" of units amounts to a preliminary estimate of the result).
3. Carrying out the mechanics of measuring (e.g., lining up the starting point on the scale with one end of the object, etc.).
4. Judging adequate level of precision.
5. Reporting the results (e.g., correct use of fractions, decimals, unit).

It is not claimed that this is a model of the psychological skills or processes employed in the act of measuring. While it is informed by, it is not meant to exhaust or replace a philosophically sophisticated theory of measurement in the natural sciences. Rather, it serves as a search template for discovering and categorizing various difficulties children have with particular measurement tasks. As the project staff worked with more and more children, directly and through the teachers, they came to understand which subskills of this model were most difficult for children who were asked to measure length. The general model was also refined and extended through experience with children. (In the particular case of length measurement, the model was extended by the analysis of the fine structure of the general subskill of scale reading.)

As a halfway station between the game and a paper and pencil test, the staff constructed and revised approximately four times a "work booklet" on length measurement. An important feature of this booklet and of the tests is that children construct the answer to the problem at hand rather than select an answer from a multiple-choice. This was done, in part, because the act of recognizing an answer is fundamentally different from the act of constructing an answer. A constructed answer permits much more useful diagnosis of the difficulties chidren are having with the measurement process. These work booklets were also used in classrooms and the results reported back by the teachers.

Only after this experience was accumulated did the staff feel ready to devise a one-page test that probed the most common difficulties with length measurement. Even at this stage they were still identifying and isolating particular difficulties that students had with length measurement as they tried to embody these is a small number of questions on the test.

Figure 1 shows (in reduction) a recent version of the resulting test. At the bottom of this test is shown (also reduced) the 10 cm cardboard ruler students used to do the test.

Of course, the test does not encompass all forms of length measurement. The distance to the sun, the circumference of the earth, and the diameter of an atom are all lengths whose measurement is important. The test is limited to the measurement, using rulers, of lengths of objects (lines, tables, pencils) comparable to the sizes of parts of the human body. This limitation involved instruments easy for children to understand and manipulate, measurements of the kind they could expect to encounter in everyday life, and a skill similar to those already being taught in schools. In fact, however, schools spend little time on instruction of measurement, despite the fact that, after computation, it is the largest single piece of all existing mathematics curricula.

As the final form of the test became clear, some of the games and activities developed early in the process were seen to be redundant or inappropriate for use in validation. For example, the Metric Measuring License (Fig. 2) required collaboration among children for its completion, and so probed considerably more than the measuring skills of the individual child. The validating activities were reduced to three: "real" measurement on classroom objects; the work booklet "What's My Line?"; and the board game "Measure and Move."

Although "real" measurements of length form an important part of validation activities, it was found to be impossible to validate the tests using "real" measurements exclusively.

Real measurements are conceptually unbalanced: They do not use in anything like equal proportions the different subskills identified as central to the length measurement process. Real measurements are also diagnostically weak: Only a long accumulation of experience allows discrimination among similar difficulties.

NAME _____ CLASS _____ DATE _____

EXAMPLE :

This line _____ Answer 10 (how many?) cm (what?)
is 10 centimeters (cm) long. (Check it with your ruler.)

a. How long is this line? —

a. Answer _____ (how many?) _____ (what?)

b. How long is this line? _____

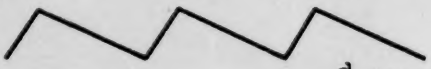
b. Answer _____ (how many?) _____ (what?)

c. How long is this line?

_____

c. Answer _____ (how many?) _____ (what?)

d. How long is this line?



d. Answer _____ (how many?) _____ (what?)

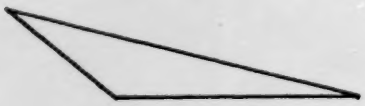
e. How long is this line? _____

e. Answer _____ (how many?) _____ (what?)

f. How long is this?

f. Answer _____ (how many?) _____ (what?)

g. How long is the longest side of this triangle?



g. Answer _____ (how many?) _____ (what?)

Copyright 1977 TOMORROW Education Development Center Inc

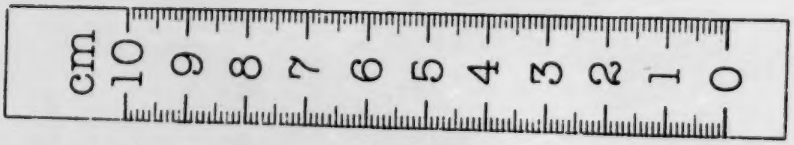


FIGURE 1.

Metric Measuring License

This license entitles _____, upon completion of the form below, to be an OFFICIAL METRIC MEASURER, authorized to measure, at any time, using any OFFICIAL METRIC UNITS (such as cm, dm, m, km, g, or kg), ANYTHING that can be measured either with OFFICIAL MEASURING TOOLS or by ESTIMATION.

height _____ cm *arm spread _____ cm* *nose to fingertip _____ cm* *elbow to fingertip _____ cm* *handspan _____ cm*

palm _____ cm *thumb _____ cm* *index finger _____ cm* *middle finger _____ cm* *ring finger _____ cm* *pinkie _____ cm*

knee to floor _____ cm *foot length _____ cm* *foot width _____ cm* *shoe length _____ cm* *shoe width _____ cm*

plain step _____ cm *giant step _____ cm* *most gigantic step _____ cm* *circumference of head _____ cm*

circumference of neck _____ cm *circumference of chest _____ cm* *circumference of wrist _____ cm* *circumference of waist _____ cm*

circumference of thigh _____ cm *circumference of ankle _____ cm* *weight _____ kg* *weight of shoe _____ g*

weight of relaxed hand _____ g

age _____ years _____ _____

signature *date*

ALL OF THESE MEASUREMENTS WILL CHANGE–KEEP YOUR METRIC MEASURING LICENSE UP TO DATE

Copyright©1974 by Education Development Center, Inc.

For example, some children round all non-integer lengths to integer values (others round to half-integer values). In order to allow discrimination of this practice, the work booklet has, as one example, four parallel lines with left ends lined up. The first and last of these lines are, for example, 5.0 centimeters and 6.0 centimeters long. The intermediate two lines are 5.2 centimeters and 5.7 centimeters long. This example is helpful in determining the child's measurement of non-integer lengths.

In using this procedure, TORQUE does not escape the use of a "derived" criterion of validity. (No one can do so, even in principle.) However, the criterion rests on a wealth of contacts with individual children and their teachers during test development and validation, on the model of measurement constructed during this process, and on several "real" measurements used directly in validation.

In passing, the central place of graphic arts in test development and validation should be mentioned. Very often children ignore printed instructions and examples and make the most amazing and unexpected interpretations of what the test is asking. Only close attention to individual children uncovers these misconceptions. Then graphic inventiveness must combine with lengthy discussions to generate the next trial version. All TORQUE tests and validation materials are "finished" in that they have the appearance of being in final published form, so children need not decode visually poor copy. A professional graphic artist is a full participant in all the project's activities.

A very gratifying aspect about this validation procedure has been the clean distinctions the tests make among children's performance on individual subskills. The overwhelming majority of children perform either consistently well or consistently poorly. This may be taken as evidence that the subskills identified are appropriate for the tasks being assessed (Schwartz & Taylor, 1978). As a consequence, clear decisions can be made

about a child's performance. Diagnoses can be made of the particular difficulties the child may be having. There is evidence that these diagnoses can lead to an effective prescription for mastery of the relevant component skills. Moreover, those who set criteria for learning do not need to ask what level of performance is good enough. For each component skill, most children either consistently perform or consistently do not perform the task.

The TORQUE Computation Package

In order to illustrate the kind of material TORQUE has prepared to assess the full range of mathematical competence in the Grades K through 6, I will describe here the instruments that have been devised to help assess children's computational competence. The two sections of the TORQUE computation package, -- the initial computation package and the extended computation package -- each probe competence with ordering and the four arithmetic operations with integers, fractions, and decimals. These materials are of two kinds:

1. Self-Checking Sheets. (See, for example, Figures 3 and 4.) On these sheets, a sequence of operations typically leads back to the orginal number or some other known number. As a result, these sheets do not require a teacher's correcting them. The children may check their own work. Since any number may be used to start, these tests can be make public. These sheets are intended to be used informally. Initially, the children supply their own numbers to start the sequence. Later, teachers may ask individual children to fill in the sheets starting with numbers that they supply. Difficulties are apparent when they occur. The children seem to be very much at ease with these sheets.

2. Computation tests printed on no-carbon-required (NCR) forms. These are formal tests, although they do not have many of the frightening

properties of other formal tests. Each sheet treats 1 computational topic. There are 10 sheets in the initial computation package and 20 sheets in the extended computation package. A sheet consists of 2 pieces of paper sealed together around all 4 edges. The child answers the questions on the first sheet. After completing the sheet, the child brings it to the teacher. The teacher separates the 2 sheets, returning the top one to the child. By folding over the right edge of the top sheet, the child can see the correct answers, and check him-or herself. The top sheet can then go home with the child. On the back sheet are the problems the child's answers, the correct answers, and diagnostic comments for the teachers. The back sheets are retained by the teacher for the child's file and for consultation with the parents. (See Fig. 5 and 6.)

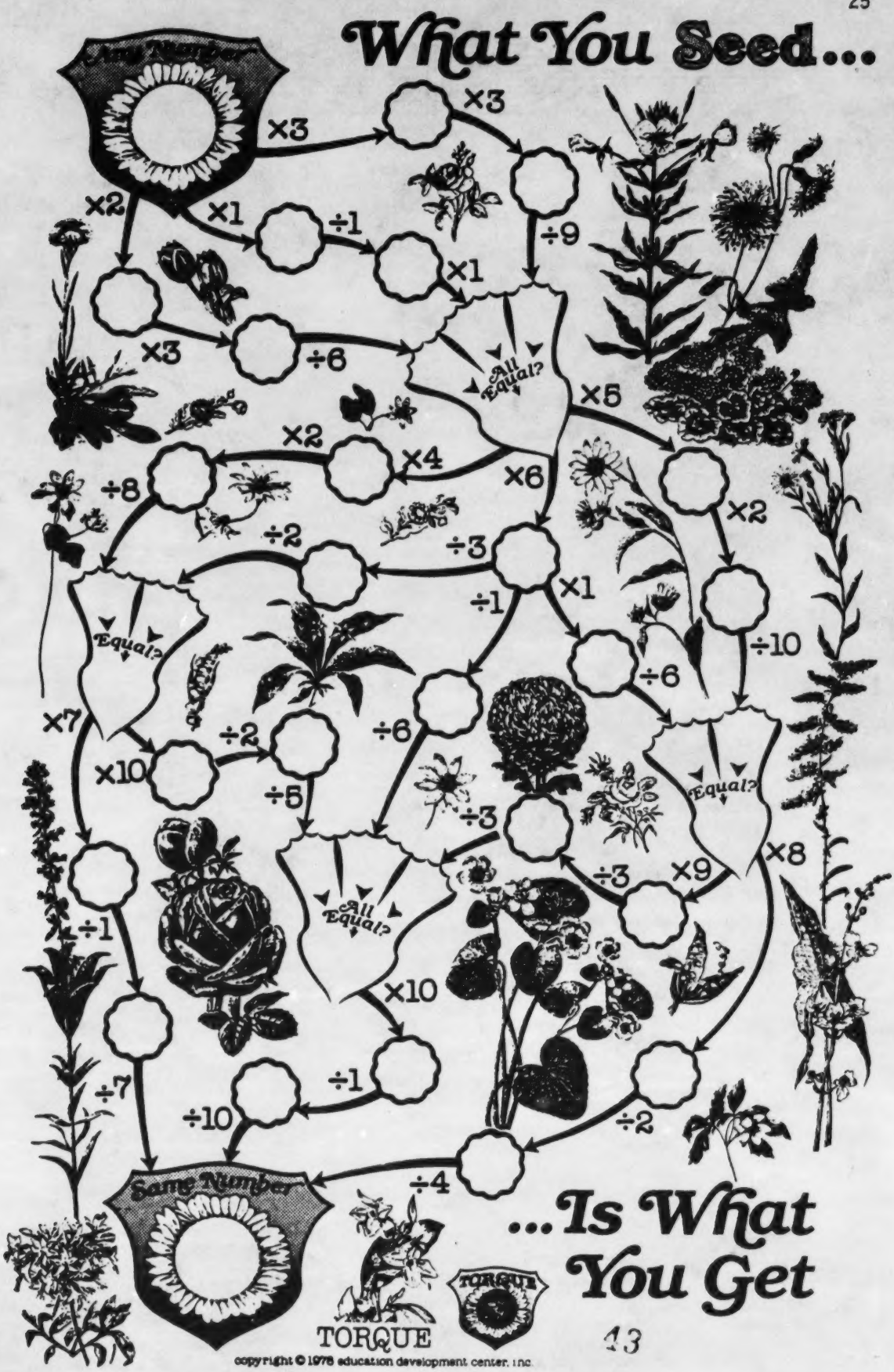
The computation tests have been validated by demonstrating that the tests are effective in diagnosing a significat majority of the computational errors made by approximately 5 thousand children and as demonstrated by the elimination of these errors after instruction.

Careful analysis of the types and frequency of errors has led to a procedure for generating a large number of different but equivalent forms of each test. Thus, the tests may be used both as worksheets and assessment devices.

The TORQUE computation package is in use on a pilot basis in lieu of traditional standardized tests, in a selected number of urban and raral school districts throughout the United States, as of September 1978.

The Setting of Performance Criteria

The development of criterion-referenced tests raises the question, Who sets the criteria? Moreover, there are obviously related questions: What topics should be taught? What level of performance is adequate? Should a given level of performance be thought of as "sufficient" for some or for all children?



copyright © 1978 education development center, inc.

Figure 3 - SELF-CHECKING SHEET

NAME _____ DATE _____

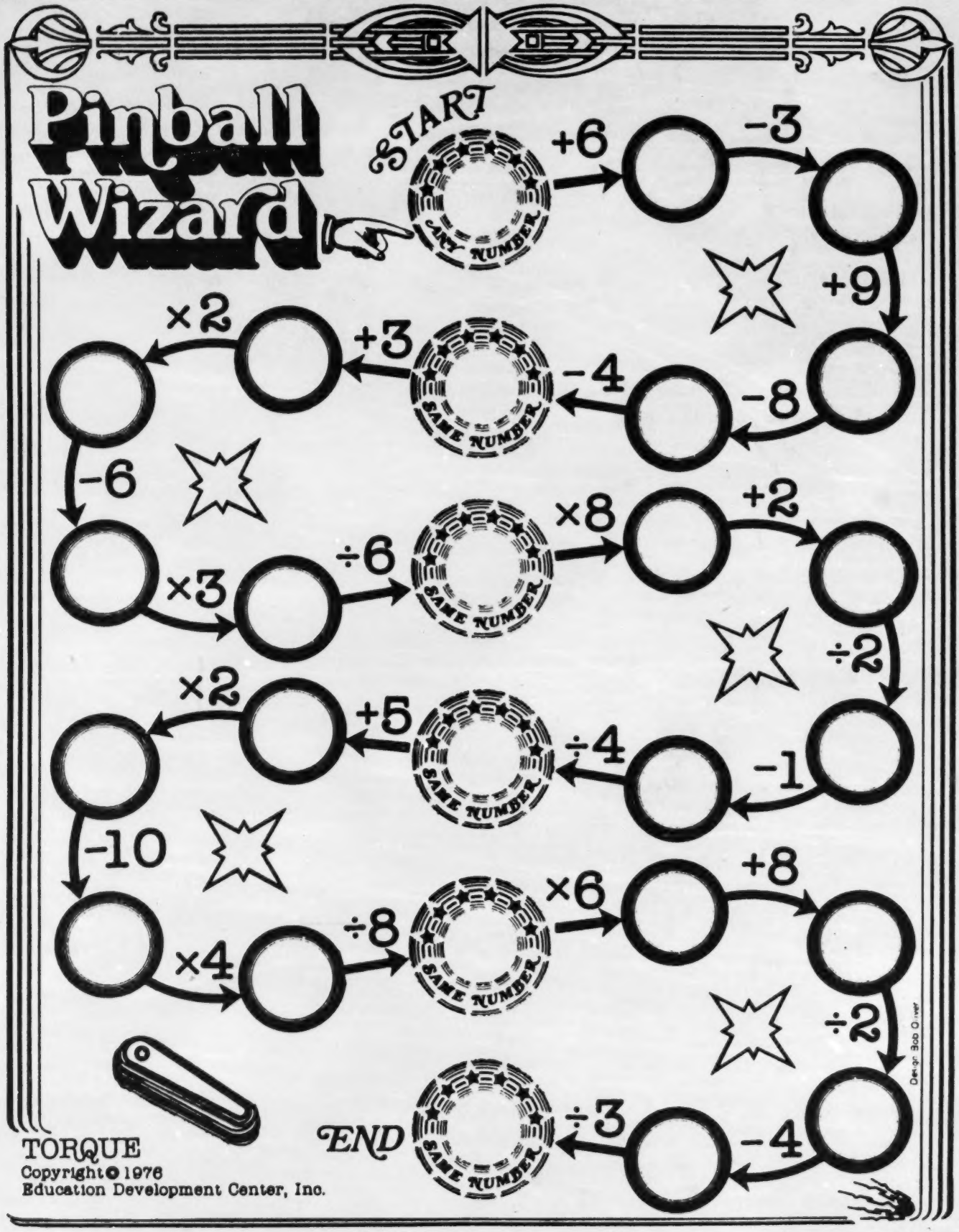


FIGURE 4 - SELF-CHECKING SHEET

NAME _____ GRADE _____ DATE _____

ADD

a. 900
+95

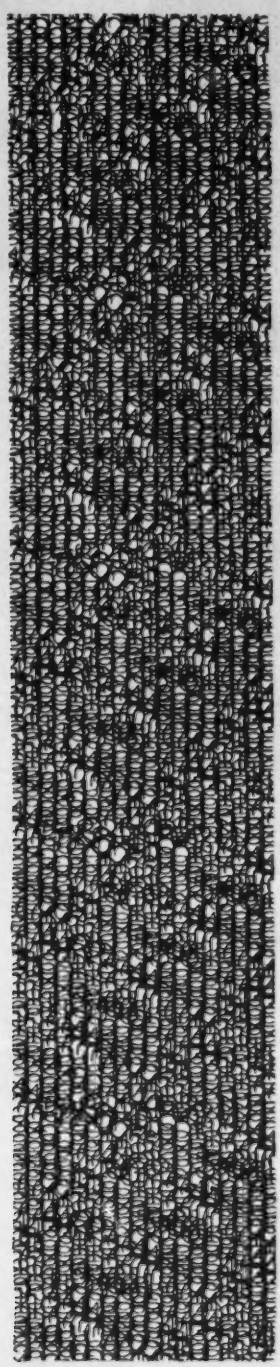
b. 904
+95

c. 734
+95

d. 916
+74

e. 936
+98

f. 795
+309



copyright © 1978 education development center, inc.

NAME _____ GRADE _____ DATE _____

SUBTRACT

Most incorrect answers arise from:

- adding instead of subtracting (add). Not a common error but one easy to identify.
- subtracting the smaller digit from the larger regardless of position (switch digits). A related error is to "bring down" to the answer the non-zero digit even if it is being subtracted from a zero digit.
- forgetting that a column has been borrowed from (forget borrow). A regrouping error.
- other errors: Occasionally students forget the borrow from the tens place and then collect all the borrows at the hundreds place.

a. 785 − 275

b. 640 − 300

c. 647 − 350

d. 856 − 508

e. 629 − 257

f. 755 − 378

g. 853 − 764

h. 735 − 48

i. 600 − 470

j. 808 − 360

k. 503 − 286

l. 700 − 358

Most Common Incorrect Answers

a. 1060—add

b. 940—add

c. 997—add
317—switch digits
397—forget borrow

d. 1364—add
362—switch digits
368—forget borrow

e. 886—add
432—switch digits
472—forget borrow

f. 1133—add
423—switch digits
487
477 } forget borrow
387

g. 1617—add
111—switch digits
99
189 } forget borrow
199

h. 783—add
713—switch digits
787
797 } forget borrow
697

i. 1070—add
270—switch digits
230—forget borrow

j. 1168—add
568—switch digits
548—forget borrow

k. 789—add
383—switch digits
327
317 } forget borrow
227

l. 1058—add
458—switch digits
452
442 } forget borrow
352

ANSWERS

a. 510

b. 340

c. 297

d. 348

e. 372

f. 377

g. 89

h. 687

i. 130

j. 448

k. 217

l. 342

IST

This test examines the skill of regrouping (borrowing) from one column to the next in the subtraction of two integers (3 digits minus 3 digits).

edc / TORQUE
copyright © 1978
education development center, inc.

7-8 form c

TORQUE has, from the outset, maintained that the ultimate responsibility for setting criteria for adequate school performance lies with society at large, informed, to be sure, by the experience of teachers and researchers. Toward that end, TORQUE has conducted two series of meetings addressed to the questions:

1. What mathematical tasks ought most children to be able to do by the end of fifth or sixth grade?

2. How can a community in general or a school in particular ascertain whether children can do those tasks?

One group, consisting of parents, teachers, principals, a math specialist, a school board member, a school psychologist, and concerned citizens held several such meetings with TORQUE staff in Washington, D.C. In a suburban Massachusetts community, a group of parents, teachers, a school administrator, and a school committee member met with TORQUE staff to deal with the same questions. In discussing computation there was little difference in the expectations of the groups.

Similarly, both groups agreed in their discussions that the goal of mathematics education, at least through junior high school, should be an easy ability to apply mathematics to common real-life situations. Typically, this application included formulating the problem, planning a solution, executing the plan (including computation), and inferring the consequences of the solution.

One outcome of these discussions has been extremely helpful to Project TORQUE. These groups help to identify which subjects in mathematics should be emphasized (and which deemphasized) in test development efforts. For example, although it was recognized from the beginning that estimation was an important skill not much taught in elementary schools, it was only after

talking with these criterion-setting groups that TORQUE fully realized how central it was to every practical mathematics task and, moreover, how much support can be generated for the inclusion of estimation in the mathematics curriculum.

Neither group progressed far enough to engage the question of what level of proficiency constitutes sufficiency. This is a key question in assessing performance of both individuals and systems. Moreover, the TORQUE staff believes that it is possible to decompose the tasks by requisite subskills sufficiently finely that performance judgments for individuals can be categorical (does the child do it or not?) rather than numerical (does the child do it some fraction, e.g. percent, of the time?). The point here is that questions of performance in this decomposition have <u>qualitative</u> rather than quantitative answers.

Setting performance levels for a school system poses a different question. Once again, it seems that one can ask a categorical question rather than a statistical one. Moreover, it is one that interacts strongly with the question of what topics should be taught. For the assessment of the performance of a system, it seems to us that a community needs to ask of any given piece of content:

-- Should (essentially) <u>all</u> the children be proficient with this content?
-- Should <u>some</u> of the children be proficient with this content?
-- Should none of the children be <u>required</u> to be proficient with this content?

At this stage Project TORQUE is addressing only the first of these questions, which is primarily germane to the vehicle rather than the tapestry subjects.

This discussion, both of individual and system assessment, and its ultimate reliance on making categorical rather than statistical judgments, may well be a special feature of the particular content matter with which TORQUE is working, i.e., mathematics. I do not take this view. I believe with others who adopt what is commonly referred to as "an information processing approach to cognitive psychology," that an analysis of a complex task, if pressed to its primitive elements, can frequently be useful in illuminating the understanding of mathematical and other problem-solving behavior in non-statistical ways.

I recognize that if school systems and communities attend to the discussions of criterion-setting groups, the result can have a substantial effect on curriculum. I believe that this is proper. If these tests grow out of a deeper understanding of the mathematical tasks and if schools attach importance to these tasks, as they do, then it is entirely appropriate that the understanding and experience gained affect curriculum content and pedagogic strategy.

IV. WHEN SHOULD ONE LOOK FOR THE EFFECTS OF INSTRUCTION?

I turn now to the second question that I was asked to address in this paper: What are the most appropriate times and strategies for measuring the effects of schooling on student learning? As before, I reject the word measuring and will address the question, What are the most appropriate times and strategies for assessing the effects of schooling on student learning?

There is a growing body of literature about human learning that points in a rather startling direction. People seem to have always assumed that growing competence at some task is accompanied by better performance at that task. One can imagine a schematic plot of performance versus competence. The traditional assumption that performance grows with increasing competence is

mirrored in a graph that climbs monotonically, better performance always being associated with a higher degree of competence. The startling character of the evidence I wish to present is that this assumption seems to be seriously flawed.

Sidney Strauss and his colleagues at the School of Education of Tel-Aviv University have been studying cognitive development in 5 to 12-year-old children in Israel for the past five or six years and have discovered a number of areas in which learning seems to go on in a less obvious and more subtle way (Strauss, In press). Specifically, they find that very young children solve certain tasks, intermediate-aged children do not solve the same tasks, while still older children solve them again. Their work has been replicated and there seems to be no doubt about the reality of the phenomenon.

The initial work of the Strauss group dealt with children's acquisition of the concept of sweetness. Sweetness is an intensive magnitude adequately defined for our purposes as being the ratio of the volume of sugar to the volume of water in a container. Four related tasks were posed. In the first task two full cups of water were sweetened--the first with two lumps of sugar, the second with one. After thoroughly stirring the sugar and allowing it to dissolve, the children were asked, "Is the sweetness in the water in these cups the same, or is one sweeter? Why?" In the second probe the first cup is filled, and the second cup is filled halfway. Each cup is then sweetened with one lump of sugar. After the sugar is dissolved, the child is asked, "Does the water in each of these cups have the same sweetness, or is one sweeter?" In the third task two cups of water again are used. The first is filled and two lumps of sugar are dissolved in it, the second is filled only halfway and one lump of sugar is dissolved in it. Once again, the question is put, "Is the sweetness of the water in each of these cups the same, or is one sweeter?" In the fourth and final task two full cups of water each have one lump of sugar dissolved in them. One of

the cups of sweetened water is then divided among two smaller cups. The three containers containing water are presented to the child who is asked if the sweetness is the same in each of the containers or if one of them is sweeter. If so, which and why?

These tasks were presented to 180 children between the ages of four and twelve. Their performance on the fourth task is most interesting. Seventy percent of the four-year-olds solved this task correctly. The percentage of children solving this task decreased steadily with age until only 10% of the eight-year-olds solved it. By age twelve, 80% of the children were once again correctly solving the task.

The acquisition of the notion of sweetness is not an esoteric rarity in this regard. Strauss has been able to collect documentation from the research literature on the occurrence of this sort of non-monotonic development in a wide range of kinds of learning; for example, in metaphor production, in language acquisition, in facial recognition, in object tracking in infants, proportional reasoning, and in the acquisition of many motor skills. But one is not obliged to turn to the research literature to corroborate the phenomenon. All of us as teachers and parents recall many occasions in which the performance of young people on tasks on which they just received instruction deteriorated as they attempted to bring into consonance newly recognized dimensions of the task with those already mastered.

Let us turn now to a consideration of the implications of this phenonmenon for assessment. What are the implications of non-monotonic development with respect to the question of optimum times to conduct assessment? The situation is interesting. Consider the following graph that depicts schematically performance as a function of competence in a skill that is characterized by non-monotonic development.

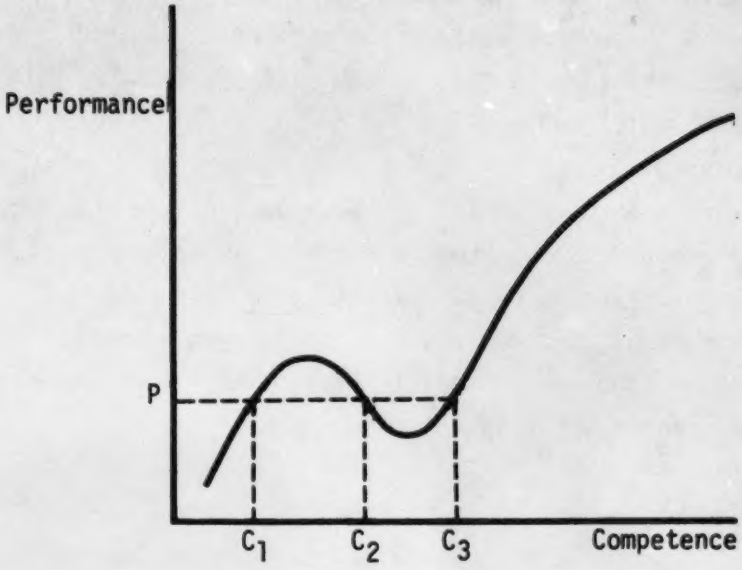


FIGURE 7

Let's suppose that in order to investigate a child's competence with respect to this skill, we have an assessment instrument that is acceptable in all respects. That is, let us assume that it is valid, that it is non-threatening, that its language is not confusing, that the child is motivated to do it, that the technology of recording an answer is not distracting, and so on. Suppose the level of the child's performance as indicated by the instrument in question is P, as shown on the graph. What inference can one make about the level of the child's competence? Clearly, from the indicated level of performance, one might infer a competence of C, or of C1, or of C2, or C3, as indicated on the graph.

The moral of the story, to me at least, is clear. No strategy of assessment that is intended to yield a "snapshot" of a child's cognitive development can, in principle, avoid this pitfall. No matter what the intended purpose of the assessment is, continuous rather than episodic assessment seems to be the only way to avoid this problem.

Thoughtful teachers and students have always understood that teaching and testing cannot be separated. A good teacher is always testing in order to find out whether he or she is being clear, or whether his or her way of presenting the subject is getting across to the students. Thoughtful students have long realized that the best tests are those that teach them something. The conclusion that we reach regarding the need for continuous rather than episodic assessment, therefore, makes it clear that a sensitive intertwining of teaching and testing is not only desirable, but logically necessary.

V. USING ALTERNATIVE ASSESSMENT APPROACHES WHAT CAN ONE LEARN ABOUT CHILDREN?

A Kind of Overview

There are many dimensions to a child's development, and academic achievement is only one of these. It is, however, that dimension of the child's growth and development that schools and educators have a special responsibility for. It would seem plausible, therefore, that schools and educators would have a particular interest in instruments that would help them discharge this responsibility in the most creative and productive manner possible.

What can teachers, principals, and superintendents learn about a child's intellectual growth and development from the child's performance on a standardized norm-referenced test? Adequate performance on such tests is

always defined in terms of the performance of others and not in terms of the task itself. As a consequence, it is difficult, even in principle, to see how diagnostic information may be derived from a child's performance on a norm-referenced test.

The multiple-choice format used in most standarized tests is a further crippling feature of these tests, when we pose the question of what can the teacher, the principal, and the superintendent learn about a child's intellecual development. The multiple-choice format does not allow for the observation of a child's formulation of a problem; it only allows the child to choose among several possible formulations that the test maker has devised. The assumption is that one of the test maker's formulations is "correct" and the others are "distractors." By not allowing the child to construct his or her own answer to the question, by forcing the child to choose among several formulations that the test maker has provided, we abandon the possibility of gaining insight into the child's own thinking about the issue. The price of such abandonment of information is high. One need not be a very subtle pedagogic theorist in order to realize that the best instruction one can provide to a child starts by capitalizing on the child's own conceptualization of the domain.

The purpose of this section, however, is not to belabor the fact that one can learn little that is useful about children from standardized tests. The purpose is rather to ask what sorts of things may be learned from alternative assessment strategies. It is to this question that I turn in the next two sections.

Comparing Assessment Strategies in a Suburban Boston Community--A Case Study

In the spring of 1977 the associate superintendent of a well-to-do suburban Boston community approached Project TORQUE. His problem, as he described it, was as follows. This system, with approximately 5 thousand children in

Grades K to 12 was unsatisfied with its mathematics program. As is frequently the case in such instances, the articulation of the reasons for that dissatisfaction was not entirely clear. The system did administer standardized tests published by one of the major test publishers at several key grade levels. These tests, he claimed, shed no light on the sources of the difficulty with the mathematics program and were of little or no use to teachers in shaping curriculum and instruction. Further, they were of no use as a basis for discussion with parents about the details of the mathematical program in the schools.

The associate superintendent had seen some descriptive material on Project TORQUE. Feeling that the TORQUE materials could more nearly satisfy the needs of this community, he approached the project staff with the request to use TORQUE materials on a pilot basis, parallel with the standardized tests that the community had used all along. In particular, he proposed that both sets of tests be administered to all fourth and sixth graders in the system. He attached particular importance to these grades because the system had several fifth and sixth grade middle schools. The fourth and sixth grades, therefore, were transition grades into and out of the middle school in that community. Anxiety about the quality of the mathematics program in the community was more nearly focussed at this level than at other levels. It was felt that this would be a good place to engage the question of alternative assessment strategies first.

There are many fascinating aspects to the entire undertaking, some of which are documented elsewhere (Cawthorne, Note 3). For our purposes here, it will suffice to focus on a small study carried out by the TORQUE staff on some of the data. In each of 17 classrooms in which both the standardized tests and the TORQUE computation materials were used, the papers of 6 children were chosen for further detailed inspection. Three of the children were those who had the highest percentile ratings in the class according to the standardized test, and the other 3 were those who had the lowest

percentile ratings according to the standardized test. For each of the 6 children, the information provided to the teacher by the standardized test-reporting form was compared with the information provided to the teacher by the TORQUE material. The contrasts were illuminating.

In my view, far and away the most important advantage the TORQUE materials offered was the ability to identify the weaknesses in the performance of the strong children and the strengths in the performances of the weak children. For example, in one classroom a child whose performance on the standardized test was in the 92nd percentile had great difficulty with computation with fractions in the TORQUE materials. The spotty nature of the coverage of fractions in the standardized test, as well as the lack of useful diagnostics, made it difficult to learn of this child's problems in this area. In the end, the TORQUE materials indicated clearly that the child's difficulty with fractions went back to a lack of grasp of the conceptual structure of fractions themselves. Under such circumstances, it is of little use to work on the child's problems in this area. In the end, the TORQUE materials indicated clearly that the child's difficulty with fractions went back to a lack of grasp of the conceptual structure of fractions themsleves. Under such circumstances, it is of little use to work on the child's computing with fractions until the basic conceptual structure of the idea is better mastered.

The TORQUE materials also proved very useful in eliciting the strengths of the children who performed poorly on the standardized test. Consider, for example, the following case. There was a child who did very well on all of the TORQUE materials but turned up in the 8th percentile on the standardized test. Of the 60 problems on the standardized test, the child had managed to complete only the first 16. Of the 16, the child completed, she had gotten 14 correct.

The teachers in this community reported they found the diagnostic information returned to them by the TORQUE materials to be of use in two ways: (a) It provided them with insights into how next to proceed with individual children who are having difficulty with various pieces of the mathematics program; and (b) as patterns of difficulty emerged among groups of children, it became a substantial aid in helping the teachers reshape group instruction so as to answer what seemed to be difficulties with their instructional program.

Teachers, however, were not the only ones who derived information from the TORQUE materials. Because the tests were printed on no-carbon-required (NCR) sheets with the correct answers on the back side of the upper sheet, the children were able to correct their own mistakes immediately upon turning in their sheets to the teacher. In a very large number of cases, the immediate quality of the feed-back was sufficient to trigger in the student a recognition of the source of error and its correction. To me, this last point is particularly interesting because it raises explicitly the notion that one of the groups that can legitimately expect to learn something from assessment instruments is the students themselves. Indeed, one might even argue that bringing people to the point of recognizing the importance of continuing self-assessment is one of the major goals of education.

A Closer Look at Learning From Alternative Assessment Instruments

When assessment instruments are designed with a deep understanding of the intellectual structure of the task and with a great deal of observation of children's behavior, it becomes possible to expand in a dramatic way the nature of what can be learned about the child's intellectual growth. Let me illustrate this point with three examples.

Counting is the name of the act we perform when we assign magnitude to a collection of descrete objects. Measuring is the name of the act we perform when we assign magnitude to a continuous quantity. Thus, we count apples but we weigh an apple. The weight of an apple can vary continuously, but one can only have an integer number of apples.

When we count, we begin with the number one; when we measure, we measure from zero. In assessing children's competence in the measurement of length, it is important that the instrument the child uses, i.e., the ruler, provide, and possibly even provoke an occasion for the child to make a distinction between an act of counting and an act of measurement. Thus, the ruler produced by Project TORQUE and shown in reduced form on the bottom of Figure 1, does not, as most rulers do, begin its scale from one end. Rather, the scale begins 1 cm. from the end. This explicity allows for the appearance of zero on the scale. Some children, whose model of the act of measuring length requires abutting the end of the ruler with the end of the length to be measured, will measure a 5 cm.-long object as being 4 cm. long. Other children whose model of the act of measuring length requires starting from 1, will measure a 5 cm.-long object as 6 cm. long.

Subtlety of content is not the only dimension along which thoughtfully designed alternatives with constructed answers allow us to learn a good deal more about a child's understanding. A child's understanding of the language of the question may differ from that of the test maker. For example, I have seen children who will not use the word <u>long</u> to describe a horizontal distance, but rather only a vertical one. Similarly, I have found odd uses of the word <u>wide</u>. At first blush, one might think that the rule governing the use of the word <u>wide</u> is rather simple: Given a roughly rectangular object, the longer dimension is its <u>length</u> and the shorter, its <u>width</u>. Anyone who thinks the issue is resolved by that might consider which dimension of his or her television screen is its width or how the dimensions of an office desk are recorded in an office-equipment catalog. There are

many more examples of the subtle interaction of the language of the task and the content of the task. Standardized-test instruments ordinarily do not make it possible for us to make this distinction.

Finally, people differ in their background and experience. A task posed in one context may offer a child no difficulty, while the "same" task posed in a different context may be a problem for the child. For example, Project TORQUE regards it important that a child be able to estimate time duration. Some children estimate time durations competently in the context of basketball games; others, in the context of bicycling; still others, in the context of chore around the house. In assessing a child's competence at the estimation of time duration, Project TORQUE allows the child to choose the context in which he or she will exhibit his or her competence at estimation. Given the pluralistic rhetoric of our society and our goal of helping each person grow in ways that are consistent with his or her idiosyncratic strengths, it is entirely proper to have pluralistic approaches to assessment. Standardized tests do not offer us that flexibility. If we continue to use them, our society will continue to deny itself access to the creative strengths of those who do not think like the test makers.

VI. USING ALTERNATIVE ASSESSMENT APPROACHES, WHAT CAN ONE LEARN ABOUT PROGRAMS?

Programs are complicated entities. It is not a simple matter to decide if a program is sucessful or, if not completely successful, to decide in what ways it can be improved. An educational program has many dimensions, including goals for individual students and goals for school systems, goals for small groups of children, and goals for teaching staffs. Elsewhere (Schwartz, in press), I have tried to analyze the situation. Let me sketch the highlights of that analysis.

In our society individual children form the core of our educational system. Our rhetoric pays them the highest homage. Within each and every educational program in our schools we are eloquent in our recitation of goals for the growth and development of the individual.

The next level of aggregation is the class. In our elementary schools, at least, classes are well defined units, each with its own teacher. Quality and success of programs may be looked at at the class level. When we do so, what can we say about the goals we have set for classes and for teacher? We tend to confuse our goals for classes and our goals for teachers, usually making the assumption that if either does or does not meet our goals, then we have prima facie evidence that the other is or is not meeting our goals.

There is a similar confusion at the next level of aggregation in the educational system. Groups of classes and their teachers constitute well-defined units called schools, each of which typically has a principal. What can we say about success of or quality of the program at the school level? Are our criteria for successful school programs identical with our criteria for successful school principals? Is an inadequate principal evidence of a failing school?

School boards that periodically wrestle with the problem of hiring, continuing, or firing a superintendent confront this confusion at a still higher level of aggregation. Are we always clear in distinguishing in our assessment between the quality and success of superintendents on the one hand and the quality and success of superintendents as individuals on the other?

Who sets all these goals and criteria? How are the decisions arrived at? What groups have input into these decisions? What instruments are available to examine the meeting of criteria? De facto, it seems to be the case that performance criteria for individual children seem to be set by the publishers of the country by the tests and texts that they produce. In

recent years several state departments of education have undertaken the setting of such criteria, usually in the form of levels to be achieved on standardized tests.

Criteria for successful teacher performance are rarely spelled out clearly by school authorities, parents, or even teachers themselves. A common practice is to assume that the achieving of children's goals is a direct indication of the achieving of teachers' goals--the class at grade level in reading means adequate teaching, and so on. Under such circumstances, the teaching dimension of a program becomes judged by an instrument that, even under the best of circumstances, is only tangentially relevant.

This tyranny of misplaced judgment becomes even more severe at the level of schools and school systems. It is clear that criteria for the success of school systems should be set by society. The schools are instruments of the society, and the formulation of the criteria of success of all society's institutions is the responsibility of the society as as whole, with the collaboration, to be sure, of the professionals the society employs. It is, however, rare indeed that a community devises procedures to shape in an active way performance criteria for schools, classes, and individuals. As a result, the responsibility is abdicated and, by default, determined by the worst of all possible instruments--standardized tests, administered to individual children. Successive layers of aggregation of test results generate increasingly meaningless, apparently precise, and utterly non-useful information.

There is a further complication in the assessment of programs. Clearly, there are things we would like to demand of all members of particular groups in the system. We would like all children to understand the concept of multiplication and to be able to recognize the appropriateness of multiplication as an analytic action in contexts that called for it. We would like all teachers to continually expand their own substantive

knowledge of the subjects they teach. On the other hand, there are things that we are not willing to demand of all children or all teachers or all schools, but of some of them. We do not insist that all children graduating from our high schools be admitted to prestigious universities, only some. We do not insist on a music program that turns each child into an enthusiast, but that the program be good enough so that some children become enthusiastic. We do not insist that all high-school teachers develop insight into the developmental psychology of the adolescent, but would regard a system in which none of the teachers had such insight as being indeed troubled. Finally, there are goals we all strive for that cannot be required of any one or of any group. We all want our fifth-grade chorus to be the best in the state. We all want the brilliant girl in the senior class to win in the math competition in the city. Striving for competitive superiority also provides goals, and we must decide how much we care about such goals.

In summary therefore, in assessing programs, we have the problem of understanding the performance of the individuals, the classes, the teachers, the schools, the principals, the school systems, and the superintendents. Moreover, there are for each group within the system goals all must meet, goals some must meet, and goals none must meet but whose attainment is viewed with great satisfaction. Trying to assess the full complexity of an educational program with instruments not adequate to assess even that for which they were designed is, to me, prima facie absurd.

A long time ago my mentor and colleague, Professor Jerrold Zacharias, taught me that there are four questions that constitute the essence of all evaluation. They are (1) What are we doing? (2) How well are we doing? (3) How do we know? (4) What should we do that is different? Clearly, a useful program assessment depends on all the actors in the educational drama being able to ask those questions of themselves. Traditional assessment makes that very difficult, if not impossible. Imaginative new alternatives

hold a great deal of promise in this regard, as I shall try to indicate in the next section, in which I will discuss recommendations for research.

VII. SOME PROMISING RESEARCH DIRECTIONS

In no particular order, I wish to sketch briefly three broad research activites that, if successful, could substantially strengthen the role that new assessment strategies could play in American education. Each of these topics is worth a far more extended discussion than is appropriate here. There is an immense amount of detail that has to filled in in each case. I believe that is done best iteratively and collaboratively with federal-support agencies committed to doing something in these areas.

A Project TORQUE for Language

The federal government should support a research and development effort at least as large as Project TORQUE to explore the development of alternative assessment strategies in language competence. Such a project should be undertaken with an explicit commitment to the development of alternative assessment instruments that satisfy all of the criteria that were discussed earlier in this paper. These include (1) instruments that have greater flexibility and accommodation to idiosyncratic learning styles; (2) instruments that are largely free of specific, cultural, ethnic, or linguistic assumptions and state explicitly those assumptions they do make; (3) to the largest extent possible, instruments that do not confound the language of the question with the content of the question; (4) instruments whose technology is convenient to the child rather than the computer; (5) instruments that are diagnostically useful to teachers and children; (6) instruments that can be made public in their entirety; (7) instruments that engage the student's full efforts with minimum anxiety and threat.

Preliminary efforts made by Project TORQUE to adapt its methods of developing and validating assessment instruments to language competence have been successful. Such efforts now need to be supported on a large enough scale so that there is no unnecessary delay in putting the instrument so developed in the hands of teachers and children.

Assessing the Tapestry Subjects

Unlike the vehicle subjects of language and mathematics, the tapestry subjects are characterized by the extraordinary diversity of their content. In science, for example, it is conceivable we might all agree with C. P. Snow that every educated person should understand the nuance and implications of the second law of thermodynamics. However, beyond that, some might hold the Krebs cycle to be non-negotiable content, while others, the theory of plate tectonics. In history, we might agree that all children in this country should understand the social and political origins of the American Revolution. Some of us would place great emphasis on the cultural history of the American people and their diverse origins. Others would argue that all children should have thorough grounding in American economic history. How does one arrive at a reasonable way of proceeding, given the legitimacy of all the different views and our supposed commitment to diversity?

I would like to see the federal government support a research and development group whose charge would be to develop a set of assessment strategies and instruments for any one of the tapestry subjects. This is something that, as a society, we must learn to do in ways that are honestly consistent with our pluralistic view of people's backgrounds, strengths, and aspirations.

Developing Action Research Paradigms

In the past it has largely been the case that educational research has been based at universities; the teachers, children, and schools researchers study

are somewhere else. Educational researchers rarely take into account teachers' views of themselves and their formulations of the situation they are in. Researchers tend to arrogate for themselves all claims to objectivity (see, for example, Campbell, 1969). By and large, teachers frequently find the results of academic research on education not very useful in their daily practice. Allow me to quote from John Elliott (Elliott, Note 4):

> The fact is that researchers' theories often cannot be made relevant to the complex practical situations in which teachers work. A professional researcher tends to take his concepts from a particular theoretical discipline of knowledge, e.g., sociology or psychology. In other words, the data he selects is guided by his interests as a psychologist or sociologist. Teachers on the other hand develop their concepts in the light of the practical problems they face in the classroom. Practical problems as opposed to purely theoretical ones are too complex to be understood from the standpoint of a single discipline. They require an eclectic utilization of ideas drawn from a variety of forms of thought. The professional researcher's theoretical problems tend to be determined by concepts locatable within a single discipline of knowledge. His research starts after the problem is defined. The teacher's practical problems, on the other hand, are determined by his experience and failure in realizing his aspirations and values. They are felt and experienced before they are understood. Once they are understood, the solution tends to be obvious. The solution to the professional researcher's problem tends to be a practically neutral statement. The solution to a practical teaching problem implies a decision about what ought to be done.
>
> Action research, as opposed to theoretically oriented research, seeks to solve practical problems by understanding them. The research process involves a dialectic between data and concepts. Action researchers begin to select data according to certain concepts, but as they study it, they realize that new concepts are required to understand the problem. As these are generated, they in turn dictate the search for more data. The understanding of the problem shifts. The new data generates further concepts which again dictates a further data search and another problem shift. The process is endless. There is no perfect understanding and no perfect solution to practical problems.

We must learn how to conduct collaborative action research involving the joint and mutually respectful efforts of academic researchers and practicing teachers.

A particularly interesting effort along these lines is currently under development in Community School District #4 in New York City. An action-research network of academic researchers and teachers in the district's rather unique science program is being formed. Such networks are clearly useful in a variety of educational settings. The urban setting, however, is, in my view, a particularly fertile one for this sort of constructive interaction. I look forward to following its progress closely. I hope the various sources of support for education research in the country do likewise.

VIII. TWO POLICY RECOMMENDATIONS

It is by now abundantly clear that there are no simple solutions to the problems of assessment. Nonetheless, we cannot avoid them. An individual who does not assess his or her own activities will lose control of his or her destiny. A society that does not pose such questions to itself may well fall prey to demagoguery. A community that has not engaged the question of what it expects from its children, its teachers, its principals, a community that is willing to accept the judgments derived from test instruments it cannot inspect, a community that is willing to reduce its richness and complexity to the simplicity of one-dimensional ordering or numerical scores, is a community that has gone far in the direction of abdicating its educational authority and responsibility.

Long-term policy recommendations tend to be platitudinous. The problems that confront us in urban education are important, difficult, and urgent. I restrict myself, therefore, to the short term and to that which is possible now, without major expenditure of resource or new ideological commitment.

There are two policy recommendations that I believe can be made that are practical and implementable immediately.

A. Make public all tests administered following their administration. As long as we think we are obliged to rely on standardized testing to generate data for forming all kinds of educational judgments, it is imperative that our society be able to inspect and debate the instruments that provide those data. Secret tests are no more acceptable in a democratic society than secret curricula. It is not sufficient to look at sample or typical questions.

There are two reasons for this insistence on making public all tests, following their administration. The first is so that public monitoring of the quality of the content can occur. Many standardized tests contain much that is wrong or ambiguous. Unless there are public mechanisms for discussion of the content of the tests, the children of our country will continue to pay the price of the test-makers' errors. The second reason for insisting that all tests be made public following their adminstration is to confront directly possible sources of bias in the test. Under current procedures, decisions about test bias can be made only after the fact and then only as may be inferred from statistical data. The public at large is never offered an opportunity to consider the question of the bias of the test items directly. Testmakers' assertions that screening committees go over the items before the fact in order to eliminate biased questions is inadequate. It is fundamentally patronizing on the part of the testing companies to say to any minority member of the American public that they, the testing companies, have selected a person to represent them and their experience as minority people in the United States.

Most publishers may object to the proposal of making all tests public following their administration on economic grounds. They may argue that the cost of such "freedom of information" cannot be borne by the industry. Our society has, in the last several years, borne the costs of a lack of freedom of information. Clearly, they are far higher than the costs incurred by making the information public. I believe that the test publishers should be willing to implement this resolution on their own, as an expression of public responsibility. If they are not, I believe that the legal grounds for exerting pressure on them via the judicial process are very persuasive.

B. Form goal-setting groups.

Groups of parents, teachers, principals, and school committee people need to meet regularly to set specific goals for children, classes, and schools. It must be clear to all participating in such discussions that the process they are engaged in is an ongoing one, and not one that results in a series of decisions and the subsequent disbanding of the group. If sucessful, such a group can contribute substantially to the easing of a major structural problem in American education, i.e., the growing sense of confrontation and alienation between school professionals and the community at large. For it is only with the blurring of the school-community boundary that we can hope to develop educational systems that are self-renewing and an organic part of their communities. In the end, though, the most desirable solution to the problem of assessment is one that has been understood for centuries, i.e., that the educated man or woman leads an examined life. We want our children, our teachers, our school administrators each to ask of themselves, both individually and collectively, what they are doing, how well are they doing it, how they know this, and what they should be doing that is different.

REFERENCE NOTES

Note 1. Schwartz, J. On inferring competence. Paper presented at the American Educational Research Association, New York City, April 1977.

Note 2. Stake, R. Personal communication, 1977.

Note 3. Cawthorne, J. Math achievement tests: What's the score? (Internal Report). Unpublished manuscript, Project TORQUE, Education Development Center, 1978.

Note 4. Elliott, J. Classroom action research network. Unpublished manuscript, Cambridge Institute of Education, UK, 1976.

REFERENCES

Buros, O. K. Fifty years in testing: Some reminiscences, criticisms and suggestions. Educational Researcher, July/August 1977, 6 (7), 9-16.

Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 4, 409-429.

Cronbach, L. Essentials of psychological testing (3rd ed.) New York: Harper and Row, 1970.

Houts, P. L. (Ed.). The myth of measurability. New York: Hart, 1977.

Levine, M. The academic achievement test: Its historical context and social functions. The American Psychologist, March 1976, 228-238.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. The American Psychologist, October 1975, 955-966.

Schwartz, J. L. The lost and found holy grail: Some reflections on accountability. Educational Technology, May 1978, 18, 42-45.

Schwartz, J. L., & Taylor, E. F. Valid assessment of complex behavior--the TORQUE approach. The Quarterly Newsletter of the Institute for Comparative Human Development, New York: The Rockefeller University, June 1978.

Strauss, S. Appearances, disappearances, non-appearances and reappearances of various behaviors: Methodological strategies of experimentation and their implications for models of development. In T. G. Bever and J. Mehler (Eds.), Dips in learning and development curves. New York: Lawrence Erlbaum Associates, in press.

OTHER TITLES IN THE SERIES

Volume 1
Compensatory Education: A Congressional Perspective
Christopher T. Cross, the U.S. House of Representatives

Volume 2
Effective Teacher-Support Systems to Improve Instructional Conditions in Schools
Joseph D. Liggins, Houston Independent School District

Volume 3
Assessment That Respects Complexity in Individuals and Programs
Judah L. Schwartz, Massachusetts Institute of Technology

Volume 4
Achievement Tests in Urban Schools
Eva Baker, University of California, Los Angeles

Volume 5
Instructional Ingredients for the Development of Beginning Reading Competence
Isabel L. Beck, University of Pittsburgh

Volume 6
A Discussion of the Literature and Issues Related to Effective Schooling
Ronald Edmonds, Harvard University and New York City Public Schools

Volume 7
Summer Drop-Off and the Effectiveness of Compensatory Instruction
Paul T. Hill, The Rand Corporation

Volume 8
The Educational Promise of Cultural Pluralism
Vera John Steiner and Larry Smith, The University of New Mexico

Volume 9
New Perspectives on School District Research and Evaluation
Catherine Lyon, University of California, Los Angeles

Volume 10
Effective Teaching and Learning in Urban Schools
Jane A. Stallings and Shirley W. Hentzell, SRI International

Volume 11
Mathematics and the Urban Child: A Review of Research
Marilyn N. Suydam, The Ohio State University

Volume 12
Basic Skills in Urban Schools: A View from the Bilingual Classroom
Concepcion M. Valadez, University of California, Los Angeles

Volume 13
School-Community Relations and Student Achievement
John E. Ingram, Jr. and Lois I. Bartels, St. Louis Public Schools and CEMREL, Inc.

Volume 14
What Do We Know About Teaching and Learning In Urban Schools: A Summary of the Conference Findings
Edmund W. Gordon, Columbia University